LA-UR-18-22621
Approved for public release; distribution is unlimited.

Title: Electric Power Outage Forecasting: Model

Author(s): Pasqualini, Donatella
Kaufeld, Kimberly Ann
Dorn, Mary Frances

Intended for: Report

Issued: 2018-03-27

Electric Power Outage Forecasting: Model

Donatella Pasqualini[1], **Kimberly Kaufeld**[1], **Mary Frances Dorn**[1]

[1] *Los Alamos National Laboratory*

Correspondence*:
PO Box 1663, MS P939, Los Alamos, NM USA 87545
dmp@lanl.gov +1 505 667 0701

1 INTRODUCTION

Electric distribution networks are the final leg of the network that moves electric power from generating stations to end users. The vast majority of the circuits in these networks are above ground, where they are exposed to direct damage from strong winds, broken tree limbs, toppled trees, and flying debris. These systems are also susceptible to winter storms. The combined stress of the weight of ice (icing), the increased wind resistance of the conductors, and broken tree limbs can damage lines, poles, and support structures.

These networks are generally operated in a radial configuration, which makes them susceptible to single-point failures. Some redundancy is built into these networks through circuit switching and alternative power supply points in the network. This redundancy can be overwhelmed by hurricanes and major ice storms that cause widespread damage. The exposure and limited redundancy of these networks makes them the primary cause of long-term electric outages following hurricanes and major ice storms.

This paper focuses on electric power outage forecasting for hurricane-force winds and winter conditions in a data-poor environment when distribution network models are not generally available. The remainder of this paper is organized as follows. Section 1 provides additional context for the problem setting and the importance of this capability development to the National Infrastructure Simulation and Analysis Center (NISAC). Section 2 describes the scope of the work. Section 3 describes the modeling approach used to forecast outages caused by both hurricane-force winds and icing conditions. Section 4 and Section 5 describe the model development, the input data, and the model results of the hurricane-force winds and the icing conditions models respectively. The report ends with a summary in Section 6.

1.1 Problem Statement

Distribution utilities use first principles–based power flow solvers on detailed models of their electric distribution network to simulate effects of local faults or other upsets to their systems. The origin of these faults is generally not specified; rather, the study goal is to design the network redundancy and operations to mitigate these assumed faults. These detailed models could be extended to evaluate or predict outages resulting from heavy, widespread damage scenarios, but several conditions would have to be met. Among these conditions are the availability of the model, the geolocation of the components in the model, and the availability of sufficient power system component metadata to enable accurate fragility estimates for hurricane-force winds and icing conditions.

Distribution network model availability is a significant challenge. As of 2017, detailed distribution network models and data are not routinely reported to any federal agency. Access to models and data is feasible only via nondisclosure agreements (NDAs) with each distribution utility, and there are several thousand distribution utilities in the continental United States (CONUS). Although collecting a few detailed models is feasible and desirable for focused studies, this approach is impractical for CONUS-scale studies or even within hurricane-prone areas because of the number of NDAs required.

Even if collecting a large number of detailed distribution models was feasible, the format of these models can be quite diverse; as such, consistent conversion from one format to another is difficult. Many distribution network databases are now fully geolocated; however, in our recent experience, component metadata are sparse or missing entirely and fragility estimates would require many assumptions about component quality and type.

1.2 Importance

The ability to predict electric outages resulting from hurricane-force winds and ice damage to electric distribution networks is key for electric power systems analysis for these extreme events. It is also a key step in the analysis of cascading failures in critical or lifeline infrastructure networks that depend directly or indirectly on electric power.

2 BOUNDING REQUIREMENTS AND LIMITATIONS

The electric power forecasting model discussed in this report is intended for use within the NISAC. Analysis scope, metrics utilized, available input data, and desired output results are subject to key bounding requirements and limitations.

2.1 Analysis Scope

The electric power outage forecasting model, associated literature review, and subsequent development of methods considers all CONUS locations. The model's primary function is to estimate the expected electric power outages from hurricane-force winds or icing conditions at a county-level spatial scale. The NISAC lacks access to detailed distribution network models, so these estimates will be made using area-based statistical models that are trained on utility-reported historical outage data and a set of publicly available regressors. Output of the predictive model will be the number of customers without electric power within each county in the CONUS.

2.2 Input Data: Scope and Limitations

The scope of the electric power outage forecasting model is the CONUS at the county level; therefore, input data to the forecasting tool should be uniformly and publicly available for all of the CONUS at that scale. The input data must be routinely maintained and updated by the original data provider or the NISAC so that the model can be run as dynamic conditions change (e.g., storm track and precipitation) and static or slowly changing input data evolve (e.g., population, soil moisture, and vegetation).

3 MODEL

Table 1 lists the input data or predictor variables. There are three types of predictors:

1. Static variables: variables that do not need to be frequently updated, such as population, elevation, land cover, tree species, and soil texture.
2. Dynamic variables: variables related to the event that need to be updated as the dynamic conditions change. For an event such as a hurricane, the sustained wind speed, maximum wind gust, and wind gust duration are dynamic variables. For an winter storm, the dynamic variables are temperature, ice thickness, snow accumulation, and wind intensity.
3. Time-dependent variables: variables calculated over a specific period of time preceding the event. The soil moisture and the standardized precipitation index (SPI) are the two time-dependent variables.

Table 1. electric Power Forecasting Model Input Data

Predictor	Source training	Source forecast	Temporal resolution	Spatial resolution
Maximum sustained wind	NOAA-Hurdat2	NHC	6 hours	County
Maximum gust wind	Estimated			County
Gust wind duration	Estimated			County
Population density	SEDAC 2010	SEDAC 2010		250m x 250m
Tree species	GECSC	GECSC		250m x 250m
Soil texture	Polaris	Polaris		250m x 250m
Land cover	NLCD2011	NLCD2011		250m x 250m
Elevation	DEM-GMTED	DEM-GMTED		250m x 250m
Soil moisture	NOAA-CPC	NOAA-CPC	Daily	10 x 10km
SPI	NOAA-NCDC	NOAA-NCDC	Daily	Point (weather stations)

For the dynamic and time-dependent variables we can find different data sources. For the electric Power Forecasting Model, the appropriate data need to provide (1) historical values over the same time period as that of the outage data we used to train the model (from October 2010 to July 2017), and (2) daily updated values to allow forecasting.

For each variable, Table 1 reports the sources used, the temporal and spatial resolution.

Surface Soil Moisture: As a support for the National Integrated Drought Information System (NIDIS), the Climate Prediction Center (CPC) provides a daily map of the surface soil moisture for the entire United States. The values estimated by the CPC soil moisture tool as part of the National Weather Service Global Forecast System is calculated using a one-layer hydrological model (Huang et al., 1996, and van den Dool et al., 2003), which calculates soil moisture, evaporation, and runoff using as forcing observed precipitation and temperature . The data provided by CPC cover the time range for which we have outage data to train the electric Power Forecasting Model.

Standardized Precipitation Index: The SPI is an index used to define and monitor drought. It is calculated as the cumulative probability of a given rainfall event occurring in a specific location. Using historical precipitation data, we estimate the probability of the precipitation being greater than or equal to the median precipitation from a predefined time scale (1, 6, and 12 months) using the approach described in (6) Positive SPI values indicate precipitations greater than median precipitation. Droughts are represented by high negative deviations. To calculate the SPI, we use the US weather station precipitation values supplied by the National Centers for Environmental Information (formerly the NOAA National Climatic Data Center [NOAA-NCDC]).

Maximum Sustained Wind: Historical data are provided by the National Hurricane Center (NHC) in the best tracks data set, HURDAT2, that is updated yearly. The tracks are provided on a 6-hour interval and include the maximum sustained wind during the period of the storm for each county affected by the storm. The NHC also forecasts storm tracks, supplying the maximum sustained wind. Using the NHC tracks and the HURDAT2 as the input, we calculate the maximum gust and its duration for each county affected by the storm using the wind model developed by (1).

Population Density: Data are provided by Ohio State University at 250 x 250 m resolution and estimated using the 2010 Socioeconomic Data and Applications Center (SEDAC) at Columbia University.

Tree Species, Land Cover, Soil Texture, and Elevation: Data are provided by Ohio State University at 250 x 250 m resolution and estimated from the Geoscience and Environmental Change Science Center (USGS), the 2010 National Land Cover Database (NLCD), the Soil Survey

Geographic Database (SSURGO), and the 2010 Global Multi-resolution Terrain Elevation Data (GMTED2010), respectively.

3.1 Resampling Predictors

With the exception of the surface soil moisture and the weather-related variables (e.g., temperature, precipitation, and snow accumulation), predictor variables are provided at high spatial resolution, finer than the county scale, which is the spatial resolution of the response data, historical outages, used to train the outage model. To resample the high-resolution predictor variables to the same scale as the response data, we use a spatial averaging weighted by the population density.

Soil moisture is provided as a GIS shapefile of polygons at a lower resolution than the county scale and its value is estimated using an average weighted by the area of the county that intersects the soil moisture polygon.

The variables that are weather-dependent, such as temperature, snow accumulation, and precipitation, are provided at the locations of the weather stations. The county value of these variables is estimated as the average of the stations in the counties. For counties that do not have weather stations, we take the average over the three weather stations spatially closest to the center of these counties.

3.2 Model Training Data

The forecasted response variable is the number of customers without electric power in each county. To train the model, we used historical outage data from the EAGLE-I database (EAGLE-I), which collects near real-time electric power outage data. The database includes outages reported by the utilities for most of the counties in the CONUS. The predictor variables are quite diverse and include hurricane gust wind speed, population, land cover type, SPI, and soil moisture index.

3.3 Model Implementation Data

To implement the model for real-time events, we require the dynamic set of predictor variables to be available and updated in real time.

3.4 Model

The storm data were fit using a random forest. Random forest is a nonparametric data mining ensemble approach developed by Breiman (2) that mitigates the tendency of regression trees to overfit. It creates many regression trees using a random subset of samples from the full training data set. Each regression tree is used to make an independent prediction of the dependent variable; these predictions are averaged to make the ensemble prediction. By randomly sampling from the training data, the regression trees are approximately uncorrelated and unbiased, which results in better aggregate performance.

The random forest model is trained using historical county-level electric power customer outage data for storm events that were taken from the EAGLE-I database. The following process was used to train and characterize the model:

1. Subdivide the original training data randomly into 10 equal-size subsets and create 10 new training sets by sequentially holding out one of the subsets
2. Train 10 random forest models, one for each of the 10 new training data sets, using standard methods in the software package R
3. Calculate the prediction error for the hold-out data for each of the 10 random forest models

The best-fitting model is the one that explains the most variability, has the lowest mean squared residuals, and has the lowest mean absolute error (MAE) and root mean square error (RMSE). The MAE is an absolute measurement of the difference between the predicted value and the true value, which is calculated by

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|$$

The RMSE measures the magnitude of error of the predicted values and variability and is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

To assess model performance of the random forest, we looked at variance explained in each model, RMSE and MAE for a subset of models using cross-validation, and partial dependence plots.

To estimate the influence of the different variables in the model, partial dependence plots were used. Partial dependence plots show the influence of a regressor on the response variable when the rest of the regressors are factored out (4). The partial dependence is calculated as

$$f_t(X_t) = \frac{1}{N}\sum_{i=1}^{N} f(X_t, x_{ic}) \qquad (1)$$

where X_t is the regressor that the partial dependence plot is being calculated for and N is the total number of observations. The plot looks at the marginal effect of X_t on the response f_t while accounting for the other marginal influences, x_{ic}, on f_t.

4 TROPICAL STORMS AND HURRICANE MODEL

The tropical storms model compares two types of storms: (1) all tropical storms and hurricanes from 2011 to 2016 and (2) hurricanes from 2011 to 2016. The purpose of the comparison was to see whether there are differences in the fit of the random forest models when taking into account all storms (low- and high-intensity storms) vs. only hurricanes (very high-intensity storms).

4.1 Data

4.1.1 Tropical Storms and Hurricanes

The data consist of 17 total storms (14 tropical storms and 3 hurricanes) from 2011 to 2016. The number of counties that had reported outages varied, ranging from 17 to 450. The total number of counties affected by outages per storm is provided in Figure 1. For Tropical Storm Hermine, 450 counties reported outages, which is more than all other storms. At the other end of the spectrum, Tropical Storm Don affected only 17 counties based on the outage data. Generally, this would imply that Don was a minor storm in comparison to Hermine. However, it is important to note that a large number of counties with reported outages does not imply that the storm (i.e., Hermine) caused the greatest number of outages. This is demonstrated in Figure 2, the log number of outages per storm. The three hurricanes, Matthew, Sandy, and Irene, caused the largest number of outages. Because the wind speeds and intensity are generally higher for these types of storms, they have a greater impact on the number of outages.

4.1.2 Eagle-I Data

The Eagle-I data were used to report the number of outages corresponding to the duration of a hurricane or tropical storm. For each county, utilities report outages every 15 minutes during the duration of the storm. In our model, we use the maximum sustained outage over a 2-hour window. We do this by first finding the minimum number of outages across a 2-hour window and then taking the maximum of the minimum outages during the entire time of the storm. To illustrate this, Figures 3 and 4 show the difference in calculation for Hurricane Matthew in Baldwin County, Georgia and Fulton County, Georgia. The red line is the 2-hour sustained minimum outages (centered) and the black line is the actual reported outage every 15 minutes. In Baldwin County, there is a large spike of reported outages around October 5 at 12:00 pm. However, this spike was not sustained over the 2 hours. Either the issue was resolved very fast or there is some indication that noise appears in the reported outages. Fulton County, Georgia's most populated county, has a different outage structure. Again there are spikes of outages, but the maximum number of sustained (2-hour) outages across the time period was around 750 reported outages. If the maximum number of outages was reported instead, the outages would be over-reported. For example, in Fulton County, 1600 outages is the maximum vs. 750 outages for the 2-hour sustained. The 1600 outages occurred for a maximum of 30 minutes in this particular county. We chose to use a 2-hour window of sustained outages to eliminate noise in the Eagle-I data and because customers that are affected for longer periods of time are of greater interest.

4.1.3 Regressors for Storms

- Latitude and longitude
- Month of the hurricane
- Population density, elevation
- SPI for months 1, 6, and 12 before a storm, matched to the starting month of a storm
- Weather variables:
  - Maximum 3-second wind speed
  - Duration of time when wind speed exceeded 20 m/sec for each county
  - Maximum sustained wind speed
  - Maximum sustained wind gusts
- Land cover (counties with any missing values removed)
- Tree species
- Soil texture
- Soil moisture (not used because of too many missing observations)

Month of the storm represents the month that the storm started. If the storm duration was over multiple months, it is denoted as the month in which the storm first hit land. This was to help account for the seasonality of the storms (early months vs. later months). The tree species are reported as proportions of the county that have a particular type of tree. The total proportion across all trees (41 tree species) sums to 1. The land cover consists of 11 different categories, including percent land developed, percent woody, percent shrubs. Similarly to tree types, the proportions across the land types sum to 1. SPI is reported for 1, 6, and 12 months before a storm and is matched to the starting month of a storm event. For instance, Hurricane Matthew started in October and the corresponding SPI values for 1, 6, and 12 prior correspond to September, March and the previous October. Latitude and longitude were included as an indicator of spatial location; generally counties further from the coast have fewer reported outages.

4.2 Model Results

Using raw outage counts, we fit three progressively simpler random forest models. In the first model, all of the regressors discussed in the previous section were included. In the second model, principal components for tree species and for land cover types were included in place of the proportional data. We chose three components for the tree species and three components for the land cover types based on scree plots and the negligible contribution of the subsequent principal components. The third model removed all regressors relating to land cover and tree species. These same three models were also fit using log outage counts and 100 plus counts with a log transformation instead of counts on the nominal scale. The results are displayed in Table 2. The table displays three different methods for assessing model performance: using raw outage counts, using a log transformation of the counts, and using a log transformation and adding 100 to the counts. The 100 plus counts were used so that the random forest weighted the higher outage counts more heavily. The model that explains the most variation is the log(100+counts) random forest model. The percent variation explained ranged from about 70–72% when all the storms were used in the analysis to about 73–74% for random forest models with only the data from the three hurricanes. Furthermore, the model with the first three principal components and the reduced model without land cover and tree species fit the best.

4.2.1 Cross-Validation Results

Cross-validation is a way to assess model performance based on minimizing the out-of-sample error. We used a 10-fold random hold-out validation test for our models. For each model, we randomly selected 10% of the data, trained the model with the remaining data, and predicted the data from the held-out sample. We calculated the RMSE and MAE for the out-of-sample predictions. This was repeated 10 times to cover the entire dataset.

Table 3 presents 10-fold cross-validation estimates of RMSE and MAE for the three models that predict outages on a log scale. In comparison, Table 4 presents 10-fold cross-validation estimates for the log(100+counts) scale. The cross-validation results reveal that the random forest models do not differ much from model to model; the differences in the RMSE and MAE are very small for both the log and log(100+counts) models. In the log model, there is a slight reduction in the RMSE and MAE for the hurricane-only data for the full model but no corresponding reduction with respect to the principal component analysis (pca) and reduced models. It is more important to note that the log(100+counts) results suggest that the all-storms model fits better when the MAEs and RMSEs are lower. In general, this implies that the random forest model with all the storm data performs about as well as the hurricane-only data random forest model, so we will proceed with results using all the storms. In the case of the models with all storms, the pca model and the reduced model perform about the same, so we chose to use the pca model to keep more of the regressors in the model. To further understand the impact of using different transformations, i.e. log vs. log(100+counts), a comparison of the bias, standard deviation, and RMSE is provided to help determine which random forest model to choose. To make comparisons across the two different transformations of the data, the RMSE was calculated as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\hat{y}_i} - 1\right)^2}$$

The idea is that if the difference between observation y and predicted value $\hat{y}$ is small, then the ratio would be close to 1, therefore not contributing much to the error. On the other hand, if the difference in values is high, then the ratio would be much greater than 1 when the actual counts are underpredicted and much lower than 1 when the actual counts are overpredicted. Table 5 shows the difference between the log10(counts) and log10(100+counts) random forest models for the following intervals: $0 - 100$, $> 100 - 1000$, $> 1000 - 10000$,$> 10000 - 100000$, and > 100000. As Table 5 shows, the bias, standard deviation, and RMSE are lower using the log10(100+counts) model for every interval except the low counts and very high counts. However, the high count category (> 100000) has only a few observations; with more observations, the bias and RMSE would tend to be lower. Additionally, although the bias is larger for the log10(100+counts) random forest model, that is a trade-off to get better predictions for outages over 100. Figure 5 shows how well the base pca random forest model with log10(100+counts) predicts the outages from all the storms. The range from 100 on fits fairly well; most of the observations are close to the line. The outages are overpredicted for the lower counts, but that is due to the transformation of log10(100+counts). The bias, standard deviation, and RMSE were also compared for each hurricane (Table 6). The takeaway from comparing the predictions using the pca random forest model with log10(100+counts) is that there are differences in the biases depending on the geographic region where the storm occurred. For example, the bias is not as high for Hurricane Matthew as for Hurricanes Sandy and Irene for outages over 1000. Both Sandy and Irene made landfall further north, whereas Matthew hit land in the south.

4.2.2 Hold-out Example Using Reduced Model with PCA

The all-storms random forest model with three principal components and using log10(100+counts) was used to predict outages for Hurricanes Sandy, Irene, and Matthew. The random forest model was trained with all the storms and then was used to predict outages for Sandy and Irene. Figures 8, 9, and 10 show the actual vs. predicted outages for

Hurricanes Sandy, Irene, and Matthew, respectively. The coast has the highest amount of outages, with areas further away having fewer outages. However, our model tends to underpredict the high outages that occur on the coast relative to other areas in the hurricane region.

4.2.3 Influence of Variables in the Prediction Model

Random forests provide a list of variable importance, that is, how much impact the regressors have in the model. The importance is calculated based on the contribution to the reduction in the out-of-sample error. The variable importance plot displayed in Figure 6 is for the final selected model, the all-storm data log10(100+counts) model with three principal components. The plot helps explain how much the MSE increases, given a variable in the model compared to the variable randomly assigned. The predictive model chosen indicates that month, longitude, population density, 1- and 6-month SPI, and maximum wind speed (sustained and gusts) contribute the most to the model. This is further demonstrated by the partial dependence plots.

Figure 7 shows partial dependence plots for four variables: maximum gust speed, population density, gust duration (wind speed above 20 m/s), and SPI for 1, 6, and 12 months. The outages on the y-axis are on the log scale. The plots show that the maximum wind speed and duration of wind above 20 m/s explain quite a bit of the variability in outages, with population density explaining the most. SPI at the 1- and 6-month intervals also explains some variability. The model suggests that wind is not the only important variable; the SPI and population density are also needed to help predict outages.

4.3 Comparison with Existing Work

The results of this analysis have aspects that are similar to those of previous publications (7; 8; 5; 11; 10), as wind speed, duration, population density, and SPI are all important indicators of the number of outages. The partial dependence plots showed that the higher the wind speed the more outages. Similarly, the higher the value of SPI (soil is more saturated based on recent precipitation) the greater the number of outages. These results are similar to previous findings. What differs in our analysis is the spatial resolution. In our analysis, we used a coarse resolution, county-level outage data, which results in large outage counts. In previous work, the spatial resolution was on the census track level or finer, in some cases on a 12,000 foot (3.66 km) $\times$ 8,000 foot (2.44 km) grid (7). The counts in the finer spatial resolution case are magnitudes smaller than those in the county-level data. In particular, the results presented in Nateghi et al. (7) have maximum values of 400 outages. The random forest model predicts well because it does not have as many large outages. In the case of large outages, their model still underpredicts the outages but is not on as large a scale as the county-level data.

Another difference is that in previous literature the utility information was often provided directly from the utilities, so information of transmission was provided in some cases. The Eagle-I data scrapes data from sites where outages are reported from utilities. It is important to note that not all utilities are reported in the Eagle-I data. For example, Vermont did not have any outages collected anywhere in the state; this is likely due to data not being accessible rather than no outages being reported.

5 WINTER STORMS MODEL

5.1 Winter Storms Data

The National Weather Service (NWS) maintains a database of 48 types of weather events having sufficient intensity to cause damage to property, lives, and commerce. Among these event types are six that the NWS identifies as winter storm events: winter storm, blizzard, heavy snow, lake-effect snow, ice storm, and sleet.

For these winter storm events, the database includes the following information at the level of NWS Public Forecast Zones:

- Source of information (e.g., emergency management personnel, law enforcement officials, newspaper, general public, trained storm spotters)
- Beginning and end dates and times for the event
- Brief narrative of the event

We faced several complications when working with this database. First, winter storm events in this database are reported at the level of NWS Forecast Zones, which may be smaller than, larger than, or the same as county boundaries. If multiple counties were matched with a zone, a new record was created for each county with the same storm information. If multiple zones with the same storm event corresponded to a single county, they were merged into one record at the county level. In a few cases, county and/or zone boundaries were redrawn during the time of interest. If the county could not be easily identified according to the current boundary lines, the record was removed.

Second, the criteria for a winter storm event to be significant vary by location, and moreover, the occurrence of different types of winter storms depends on regional climatology. (NOAA National Severe Storms Laboratory) In fact, the records included in this database are those that meet or exceed the *locally* or *regionally* defined warning criteria for the respective event types. For example, the criteria for a "heavy snow" warning may require 6 inches or more in 24 hours in one zone and 8 inches in 12 hours in a different zone. Furthermore, the different types of storms are not necessarily distinct. For example, a "winter storm" is a winter weather event that has more than one significant hazard (e.g., snow and ice) and meets the regional criteria for at least one precipitation element, and a "blizzard" is a type of winter storm that further satisfies conditions regarding sustained wind speeds and reduced visibility due to snow. The number of events of each type (at the county level) from January 2011 to February 2017 is shown in Table 11. Because of the ambiguity in the definitions for the different storm types and the relatively small number of certain events, such as ice storms, we chose to combine all of the winter storm types.

Finally, unlike hurricanes, winter storm events are not well defined—we have neither a list of dates for historical storm events nor a simple criterion to determine the locations affected by a storm. How can we track the movements of a storm system across space and time? These records are cleaned by the regional NWS offices, so there is no global view (across states or regions) of these storms. When two neighboring counties experience the same winter weather conditions, are they being affected by the same storm? Most likely, but how do we determine which counties are affected and the time span for such a storm? In this database, some counties had multiple overlapping events of different types. There were many cases where a county experienced more than one "event" of the same type in short sequence. Are these part of the same storm? Some of the answer may lie in the narrative provided by the NWS, but the information included in this narrative is not uniform in its level of detail or its regional specificity. In this work, we take into account only the latter (temporal) question. Because historical weather data (such as precipitation) are available only on a daily scale, we combined storm events in the same county that occurred sequentially with at most 24 hours between the end of the first storm and the start of the second. Because we do not distinguish between different types of storms in the present analysis, we merged any events that occurred simultaneously by taking the earliest start time and the latest end time for the duration. Figure 12 shows the duration for the resulting winter storm events. Most of the storms lasted only 1–2 days, but a few storms lasted as long as 11 days. The total number of winter storm events recorded in each county from January 2011 to February 2017 is shown in Figure 13. It is clear that nearly every county in the CONUS is affected by winter storms.

5.2 Eagle-I Outage Data

First, we identified the 15-minute reported outage numbers corresponding to the duration of a storm event. From these outage numbers, we calculated the maximum 2-hour sustained outage for that event using the same method as for tropical storms and hurricanes. Figures 14 and 15 illustrate this calculation for a winter storm event in Cook County, Illinois and Peoria County, Illinois. The example of Peoria County shows that there are many time points for which there is no reported outage number. If outage records were not available to account for at least a 2-hour window, the storm was removed from our dataset. Among the storms for which outage information did not account for a 2-hour window were 86 storms that lasted less than 2 hours. It is important to note that a lack of outage information does not necessarily mean that there were no outages at that time; it simply means that we do not know how many (if any) customers were affected. Unfortunately, this process removed a large number of winter storms from consideration, particularly in the mountain states and in parts of the Midwest. This significant loss of information can be seen in Figure 16, which shows the number of winter storm events in each county after merging with the outage data.

Figure 17 displays the maximum 2-hour sustained outage during a winter storm event for each county. For counties that had multiple winter storm events during the 2011–2017 period, we show the worst case (i.e., the winter storm event associated with the largest number of outages).

5.3 Regressors

- Latitude and longitude
- Week of the starting date of the storm, duration of the storm
- Population density, elevation
- SPI for past 1, 6, and 12 months, matched to the starting month of the storm (counties with any missing values removed)
- Weather variables
  - Daily precipitation: aggregated over the duration of the storm, if applicable
  - Daily snow accumulation: aggregated over the duration of the storm, if applicable
  - Daily minimum temperature: calculated maximum and minimum value over the duration of the storm, if applicable
  - Daily maximum temperature: calculated maximum and minimum value over the duration of the storm, if applicable
  - Daily fastest 5-second wind speed: calculated maximum value over the duration of the storm, if applicable
- Land cover (counties with any missing values removed)
- Tree species
- Soil texture (not used because it was available only for states affected by hurricanes)
- Soil moisture (not used)

We included the week of the year when the winter storm event began to account for seasonality, i.e., the differing impact of a storm that occurs early in the season and a storm that occurs in the middle of winter. We additionally included the storm duration to capture potential differences between a localized storm event lasting only a few hours and a large storm system affecting a larger region over the course of a week or more.

For winter storm events spanning multiple days, daily precipitation and daily snowfall were aggregated over the relevant days of the storm to create regressors representing the accumulated precipitation and the accumulated snow for the duration of the storm. For the daily minimum and daily maximum temperature, we calculated the overall minimum and maximum extreme daily temperatures over the duration of the storm. For wind, we included the maximum value over the duration of the storm.

The other regressors included in the winter storm models were SPI for past 1, 6, and 12 months, latitude and longitude of the county centroid, elevation, population density, land cover types, and tree species.

Some of these regressors were available for only a limited number of counties or for limited periods of time. Figure 18 shows the shrinking number of winter storm events as we sequentially added the regressors and removed any events for which there were missing values for the corresponding regressors. After including all regressors, 3245 historical winter storm events remained on which to train our models. However, these events represent storm events from only 400 unique counties. As seen in Figure 19, these storms offer very sparse coverage of the nation. In this map, the outage count displayed is once again the largest sustained outage number for that county.

5.4 Results

The winter storms have a large spread of outages, similar to the tropical storm and hurricane data. Although no zero outages were reported, many counties reported a single outage. There are storms associated with very high outage counts, but Figure 20 shows that particularly damaging storms resulting in outages for over 100,000 customers make up a very small portion of significant winter storms. To help adjust for the high counts, we worked with outages on a log scale, similar to the tropical storm and hurricane data.

Using raw outage counts, we fit three progressively simpler random forest models. In the first model, all of the regressors discussed in the previous section were included. In the second model, principal components for tree species and for land cover types were included in place of the proportional data. For each group of regressors, the first five principal components were selected based on the negligible marginal contribution of the subsequent principal components. The third model removed all regressors relating to land cover and tree species. These same three models were also fit using log outage counts instead of counts on the nominal scale. The results are displayed in Table 7. As in the models for tropical storms and hurricanes, the models that predict outages on a log scale explain a greater amount of the variation in the data than models on the nominal scale. Furthermore, the simplest model appears to perform better. However, none of these models appear to be capable of predicting the outages with high accuracy.

Table 7 presents 10-fold cross-validation estimates of RMSE and MAE for the three models that predict outages on a log scale. We now look at the simplest of these models, denoted as “Reduced”, in more detail.

We investigated the forecasting ability of this model by holding out 10% of storm events for out-of-sample prediction. The true outages are plotted in Figure 23, and the predicted outages are plotted in Figure 24. The model clearly underpredicted the large outage events and overpredicted the small outage events.

The variable importance plot in Figure 21 shows that the accumulated precipitation, population density, maximum minimum temperature, and longitude contribute the most to the reduced model, with SPI at 1, 6, and 12 months not contributing as much. When taking a closer look at the accumulated precipitation from the partial dependence plots (Figure 22), precipitation contributes more weight than snow accumulation. The precipitation explains quite a bit of the variability in outages. The daily minimum temperatures, both the maximum and the minimum, have a similar pattern: the outages increase the greater the values. This behavior is consistent for the other weather variables, wind gust and daily maximum temperature. The model suggests that accumulated precipitation is important but that snow accumulation does not contribute as much to the model.

6 SUMMARY

Overall, the random forest model for tropical storms and hurricanes underperforms compared to data that is on a finer scale from previous literature. The magnitude of the data (hundred thousands vs. thousands) makes it harder for the model to predict large outages. The random forest models examined in this analysis failed to predict the high outage counts, as can be seen in the plot of predicted outages and true outages for all counties in the training dataset displayed in Figure 25. Further analysis of the outage data and predictors or a finer spatial resolution and a different model approach (one that can account for zeros in the model) may help predict hurricanes. An example of a model that would account for excess zeros is a zero-inflated model. In the case of the large outage values, extreme value analysis is an approach to identify large events such as high outages from hurricanes.

1 FIGURES

1.1 Tropical Storms and Hurricanes

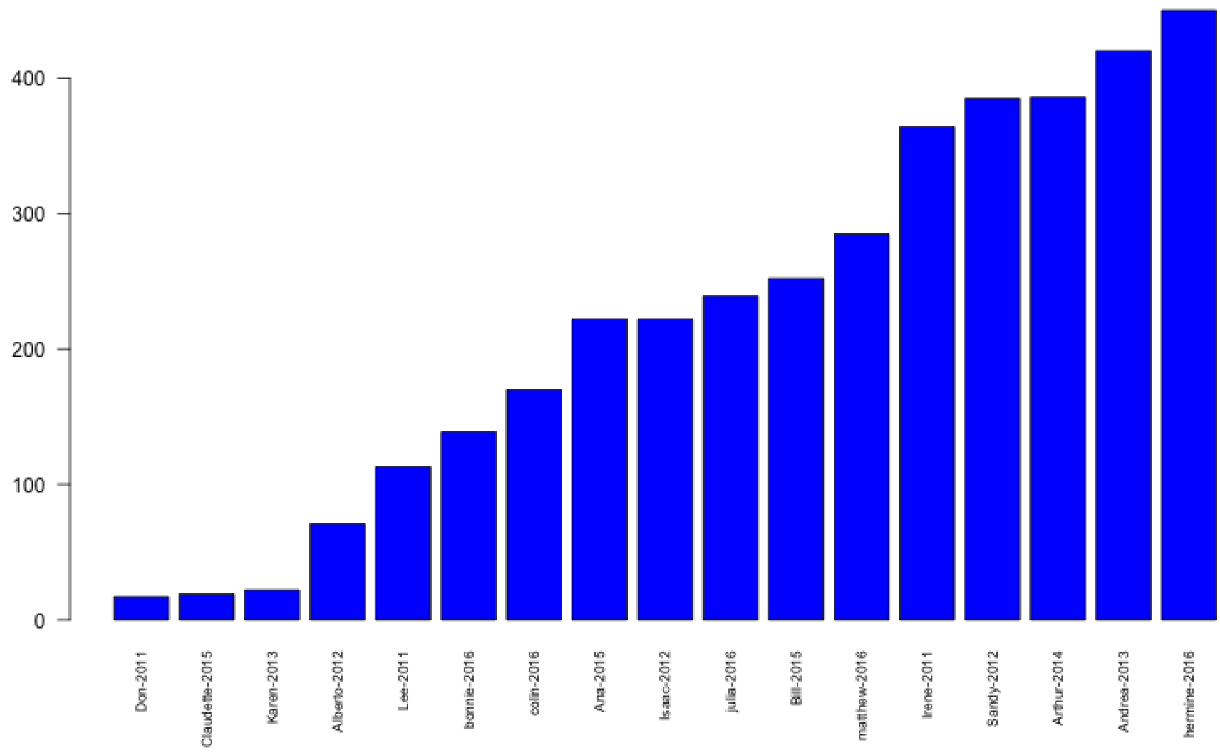

Figure 1. Number of counties per storm with reported outages

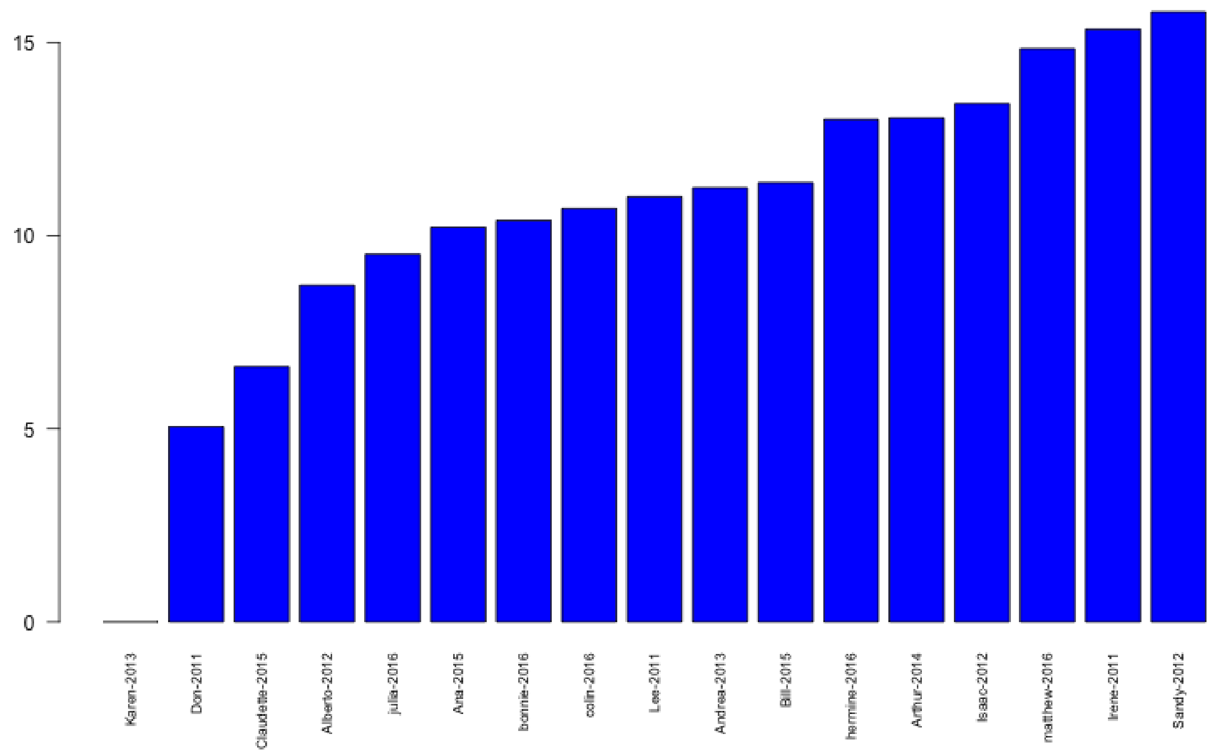

Figure 2. The log number of outages per storm

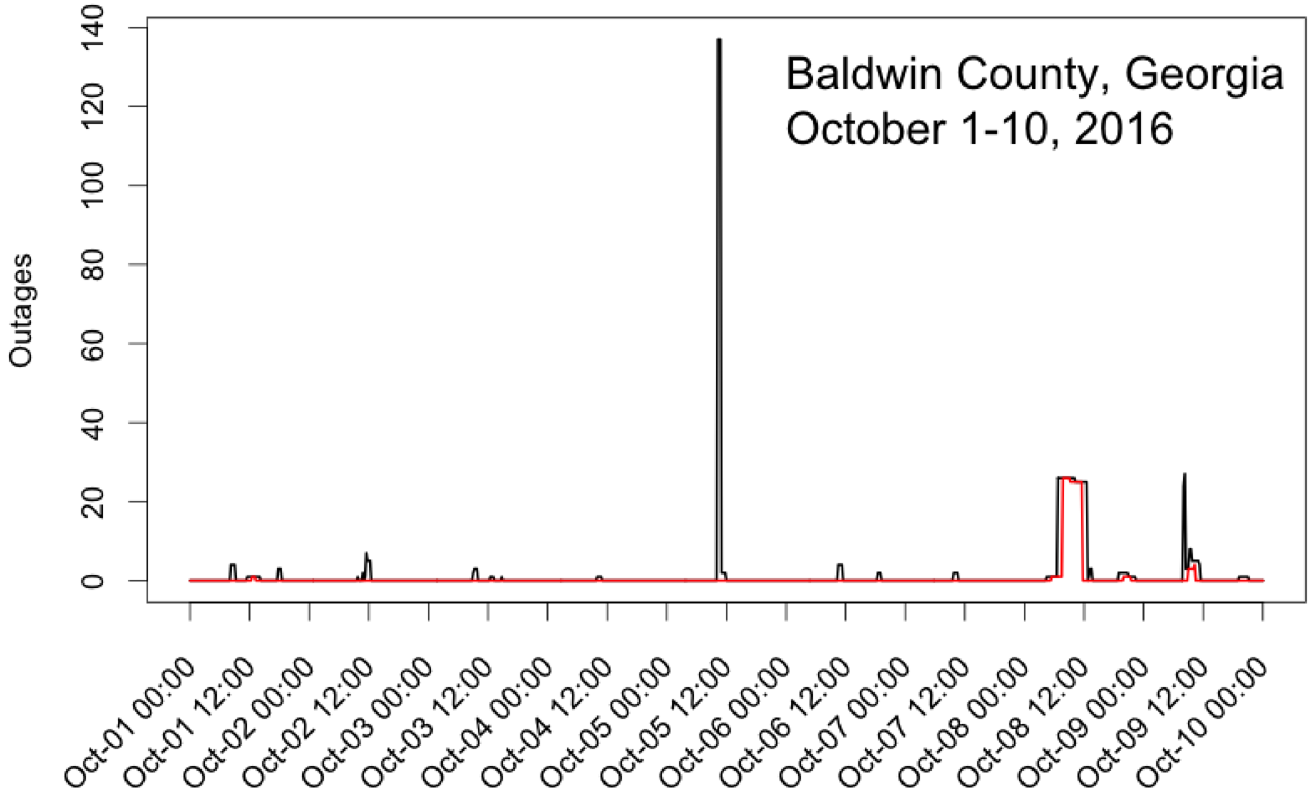

Figure 3. Examples from Hurricane Matthew to illustrate 2-hour sustained outage idea (Baldwin County, Georgia)

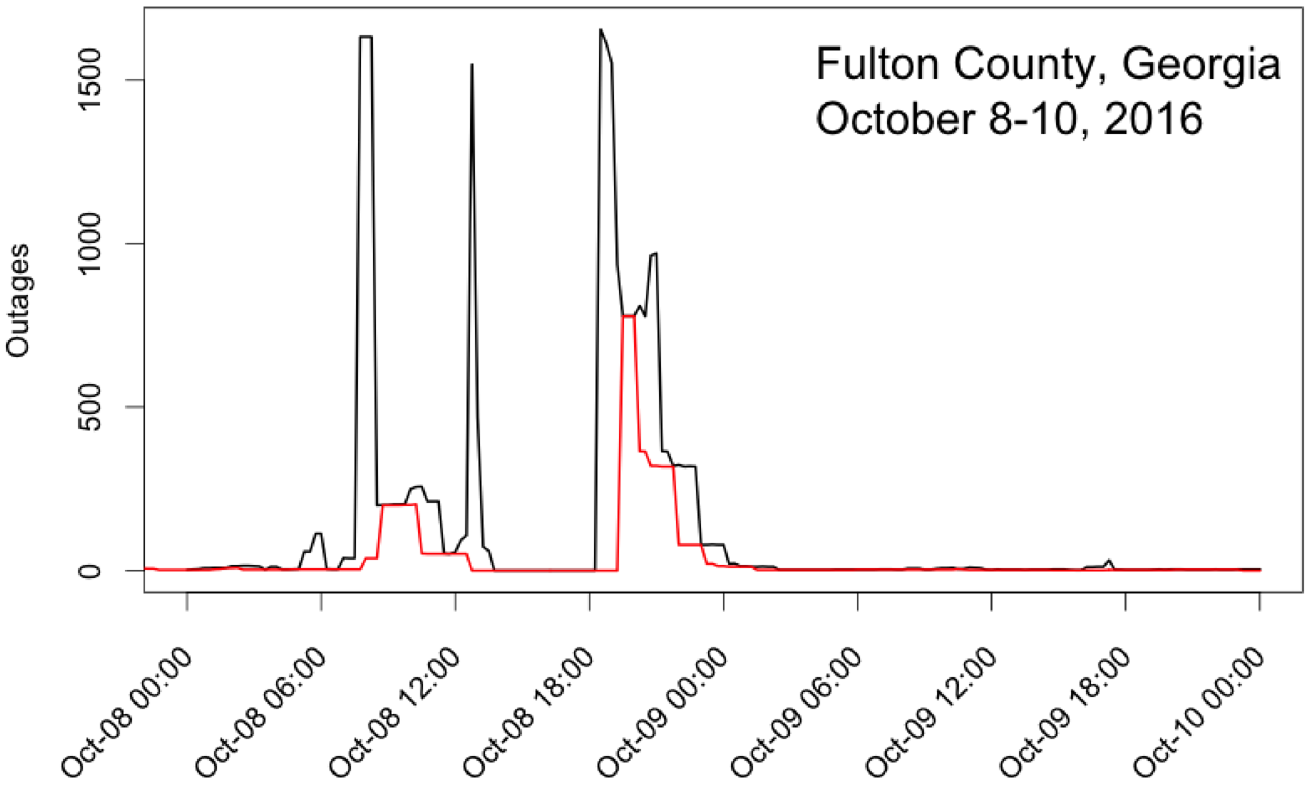

Figure 4. Examples from Hurricane Matthew to illustrate 2-hour sustained outage idea (Fulton County, Georgia)

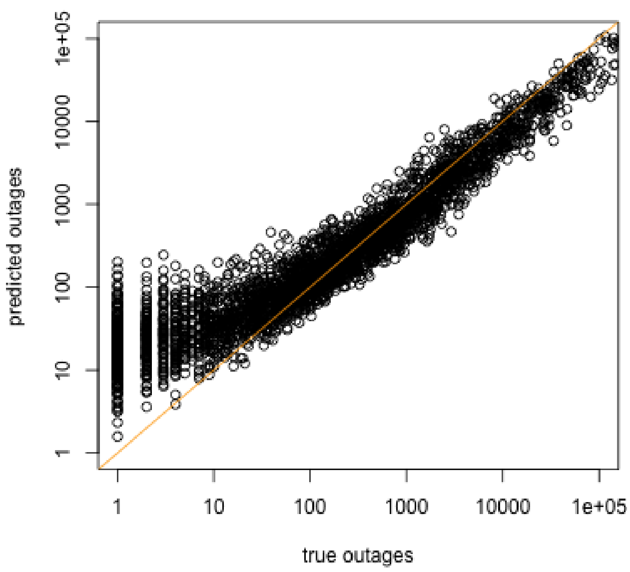

Figure 5. Alternative to the two maps: Predicted vs. true outages for predicting all storm outages

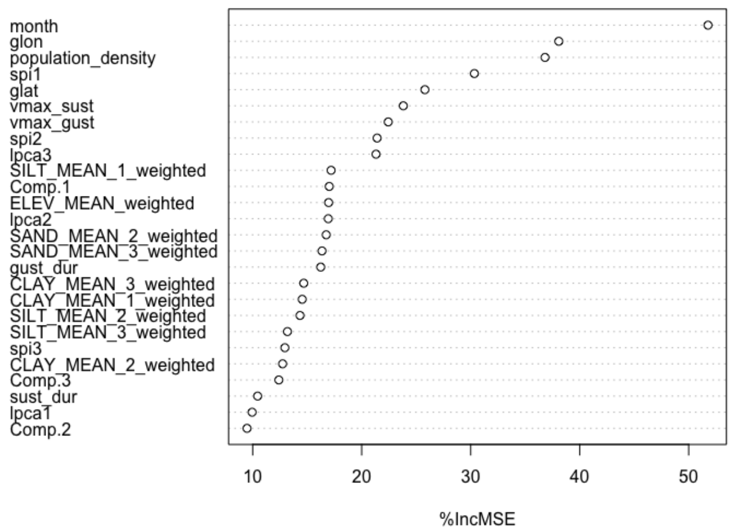

Figure 6. Variable importance plot—log scale pca reduced model

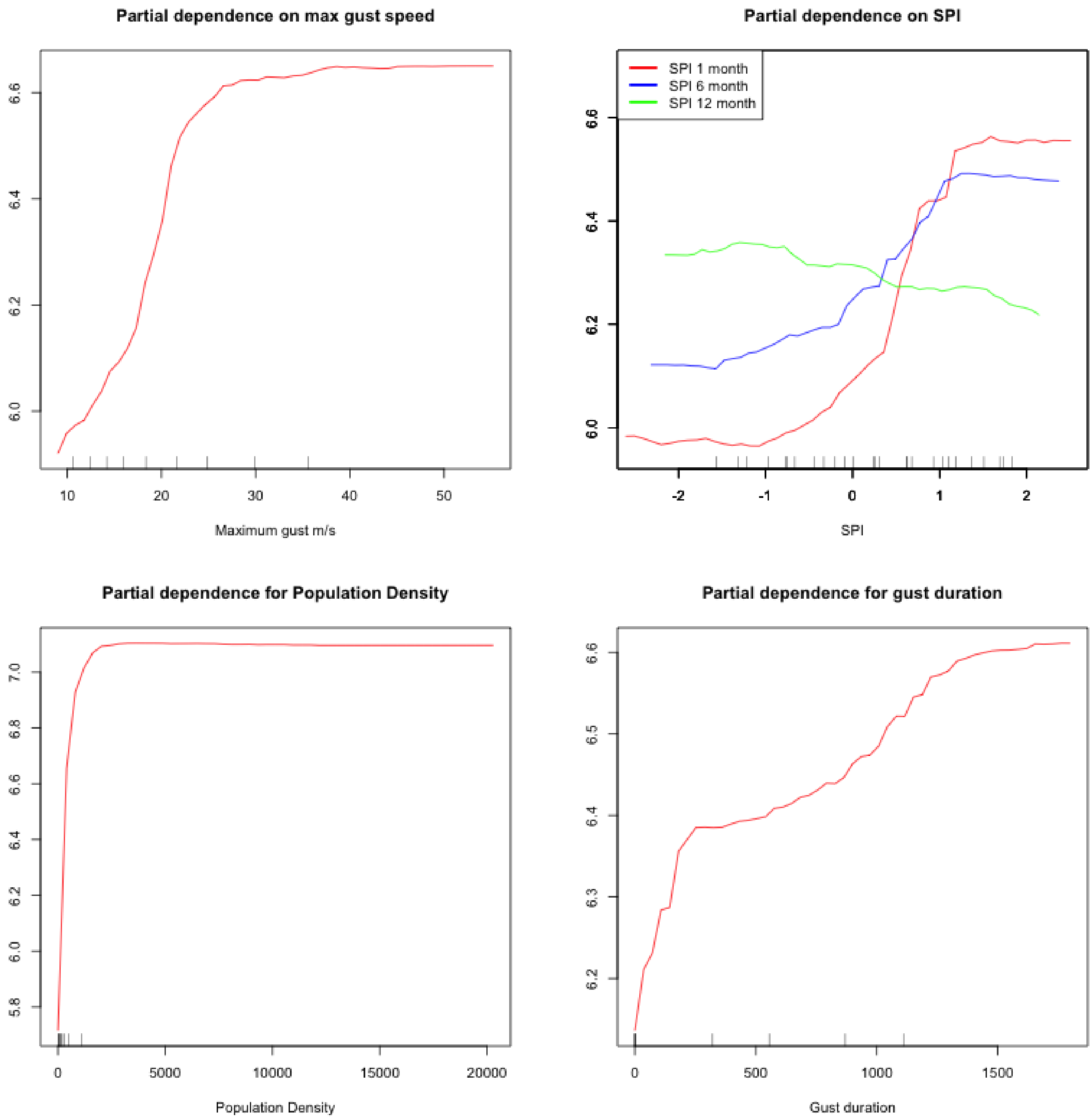

Figure 7. Partial dependence plots—log scale pca reduced model

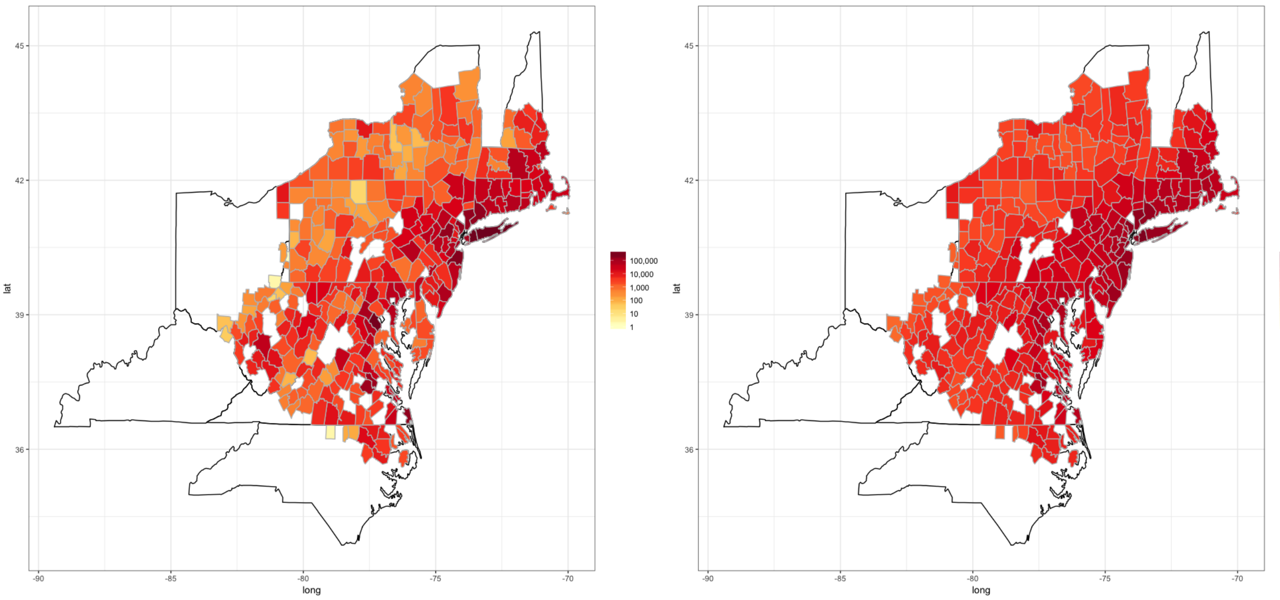

Figure 8. Actual outages (left) vs. predicted outages (right) for Hurricane Sandy

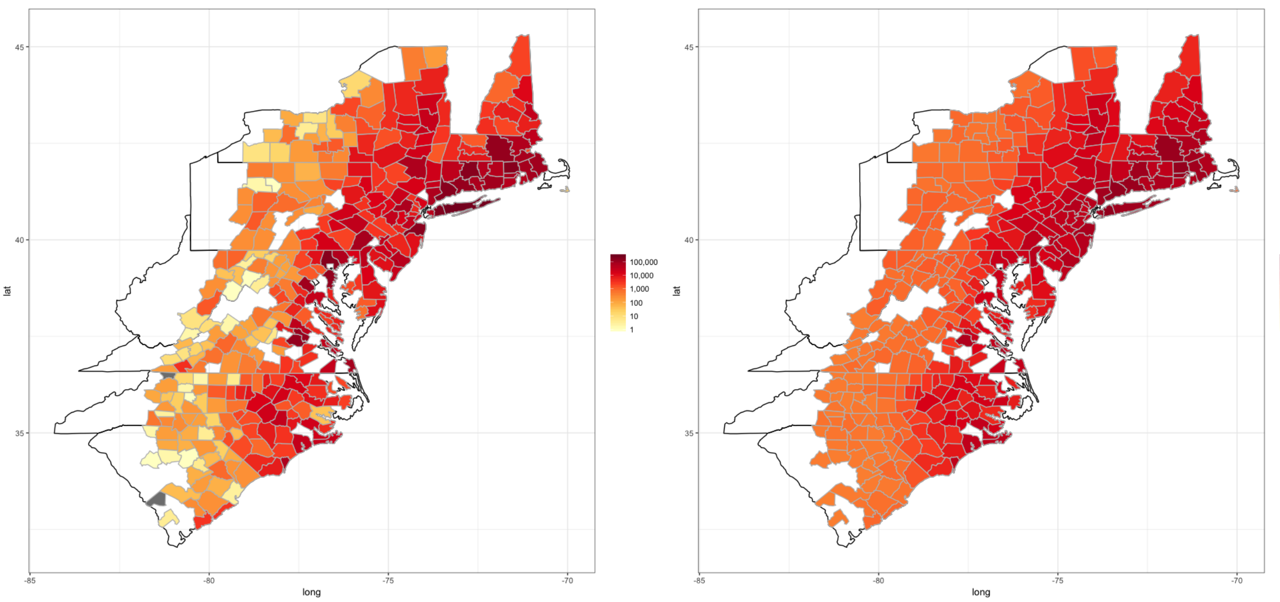

Figure 9. Actual outages (left) vs. predicted outages (right) for Hurricane Irene

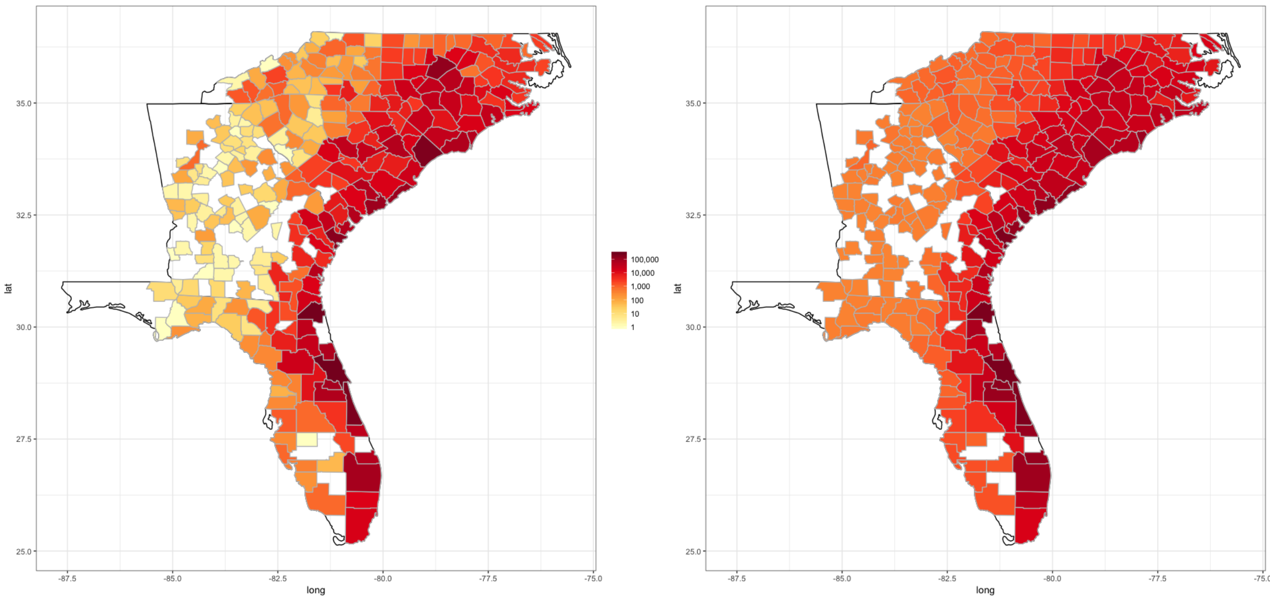

Figure 10. Actual outages (left) vs. predicted outages (right) for Hurricane Matthew

1.2 Winter Storms

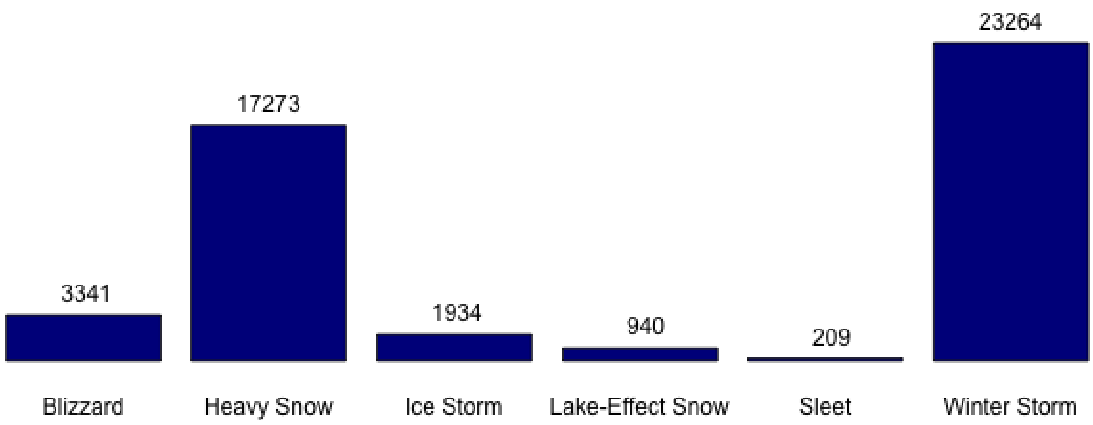

Figure 11. Number of winter storm events by type

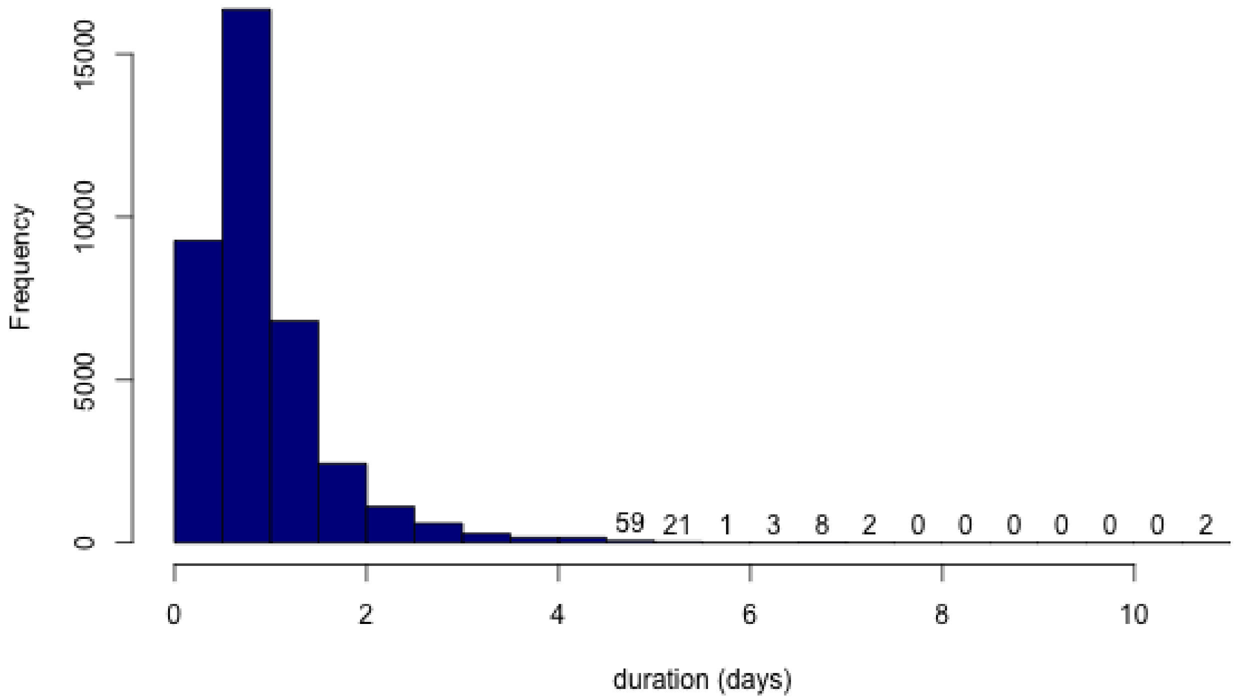

Figure 12. Duration of winter storm events

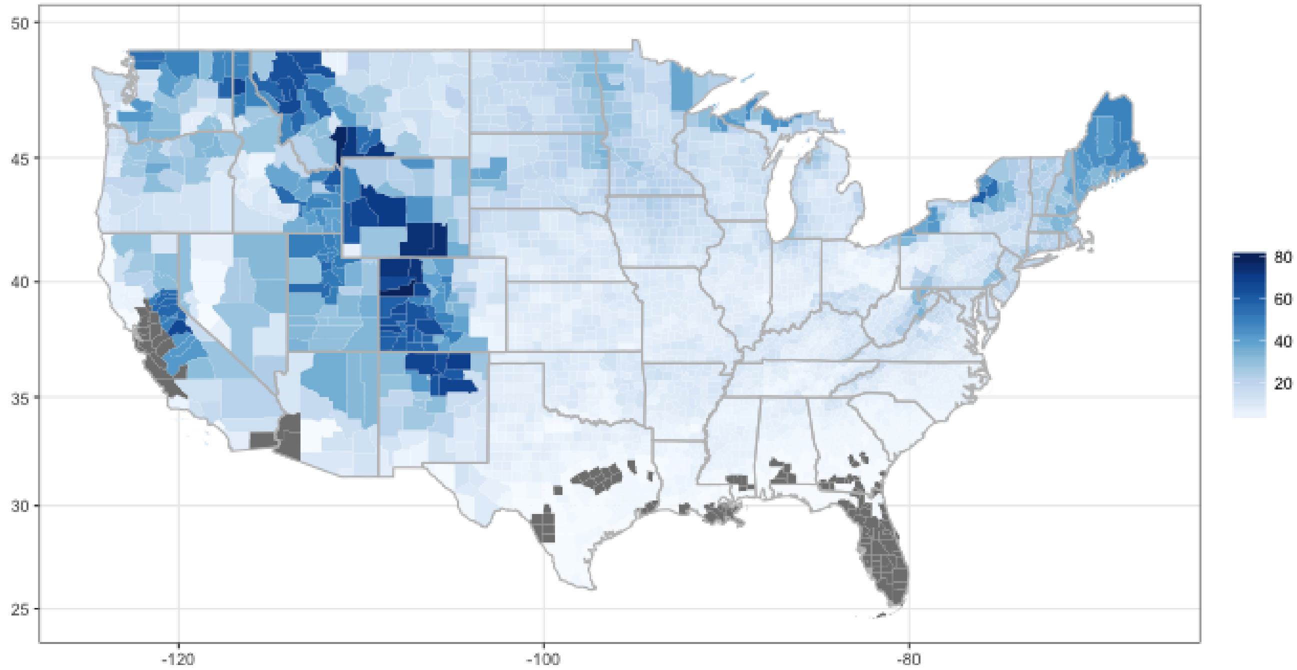

Figure 13. Frequency of winter storm events in each county from January 2011 to February 2017

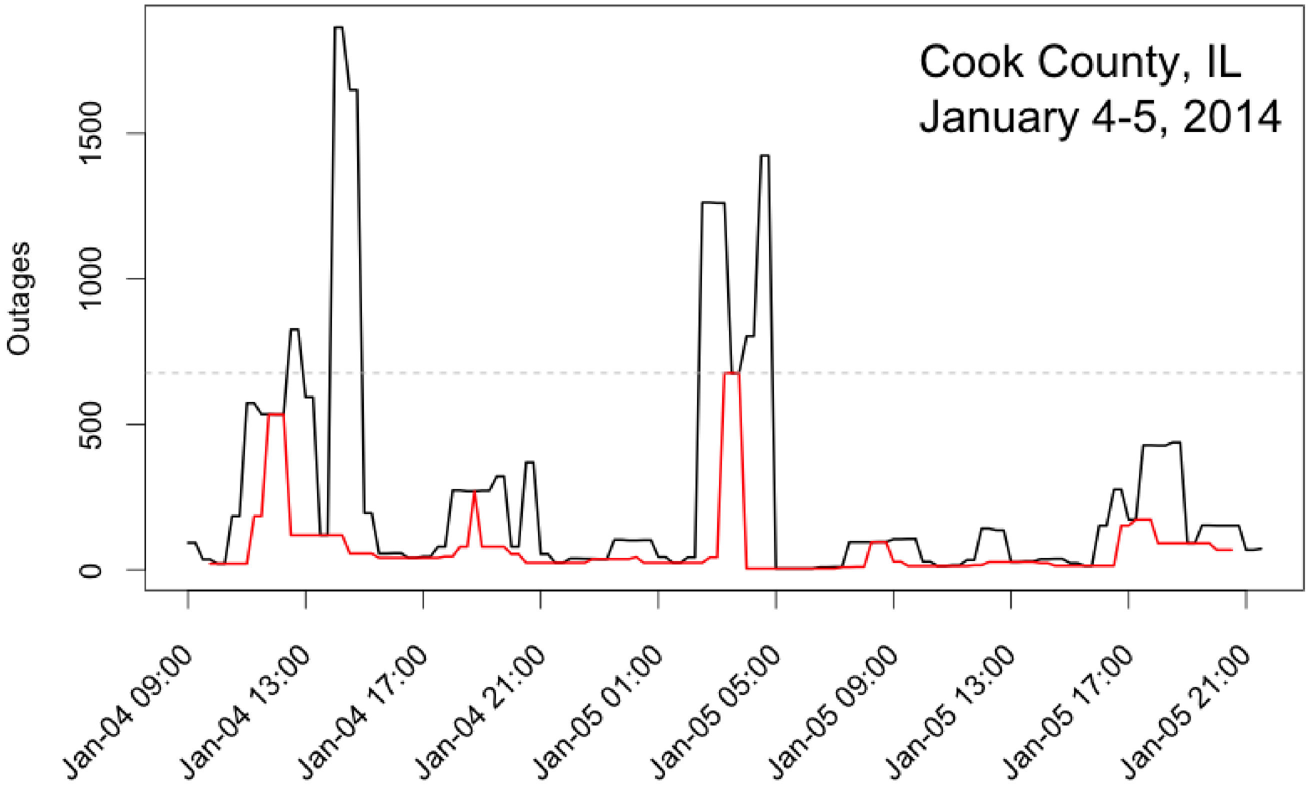

Figure 14. Examples of winter storm events to illustrate 2-hour sustained outage idea (Cook County, Illinois)

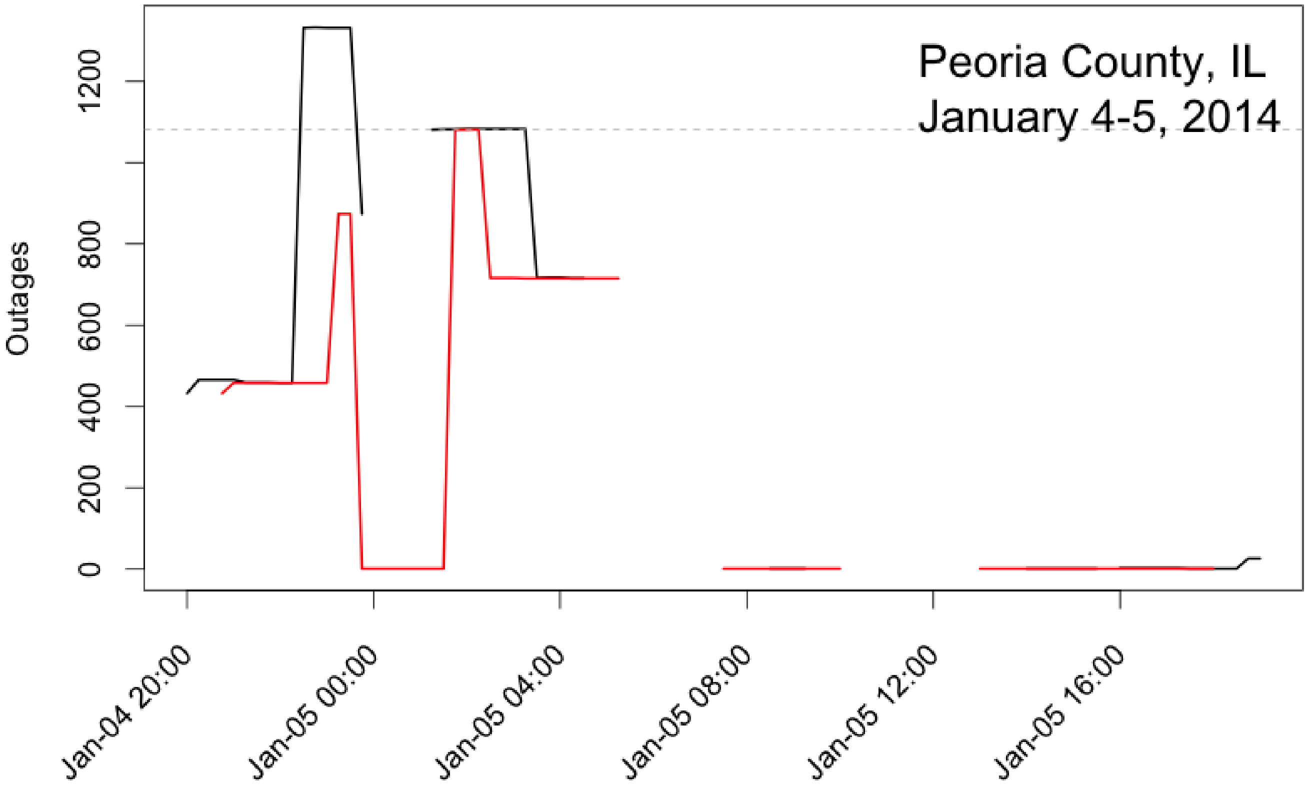

Figure 15. Examples of winter storm events to illustrate 2-hour sustained outage idea (Peoria County, Illinois)

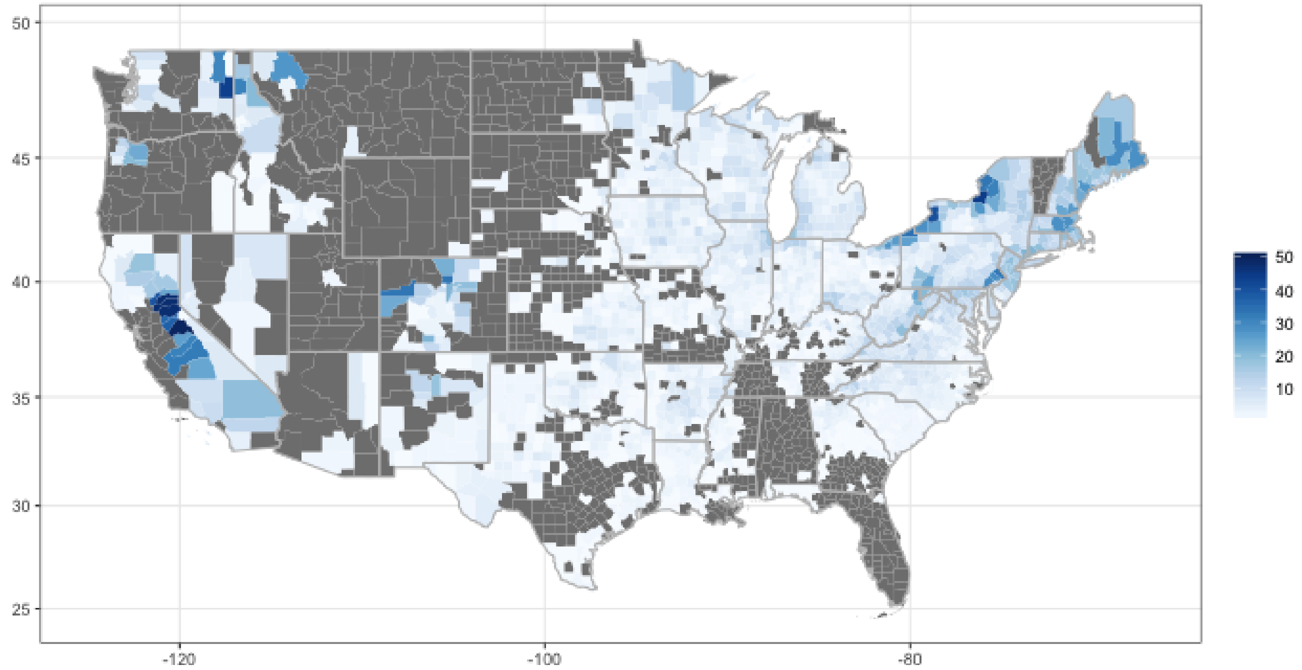

Figure 16. Frequency of winter storm events having sustained outages from January 2011 to February 2017

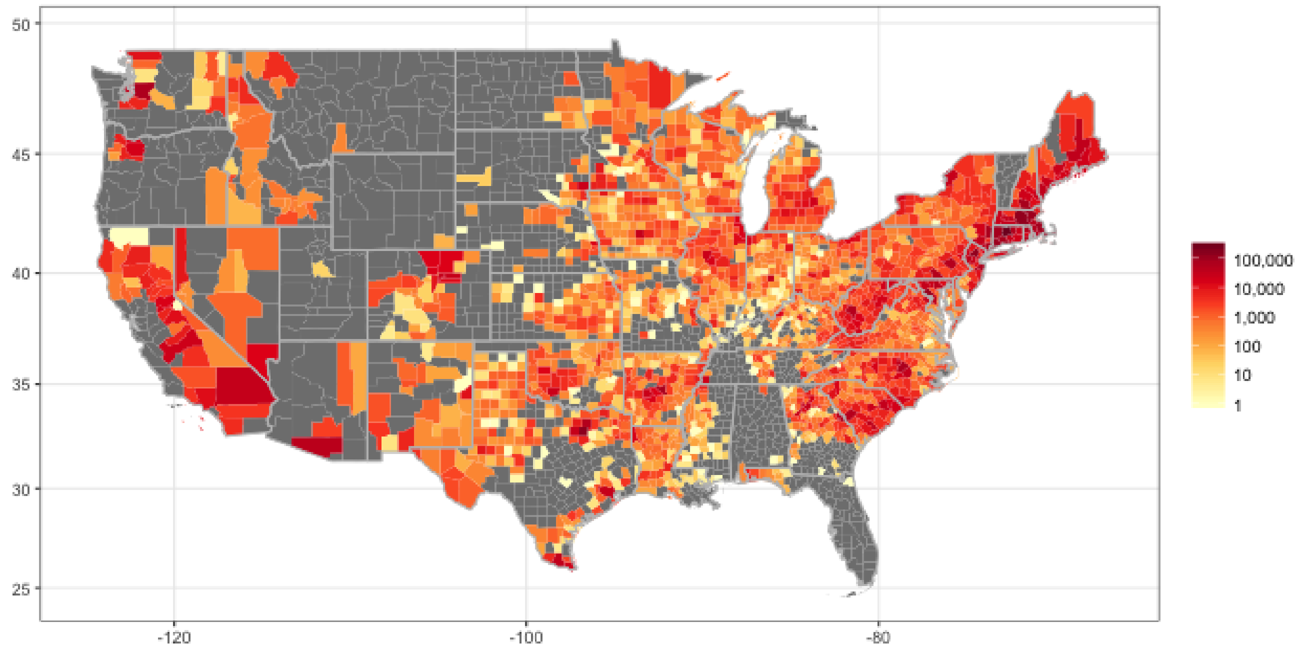

Figure 17. Worst-case scenario: the number of outages during the most significant storm event in each county from January 2011 to February 2017 (with and without regressors)

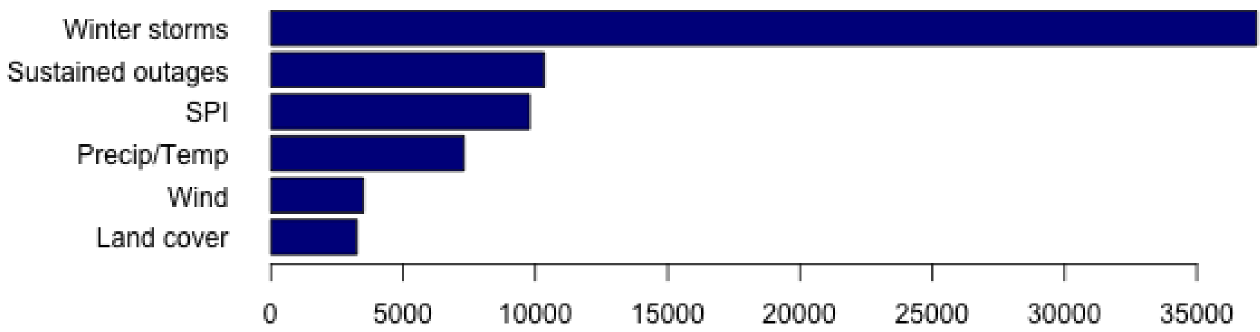

Figure 18. Shrinking data: the number of storm events in the training dataset after sequentially merging with regressors

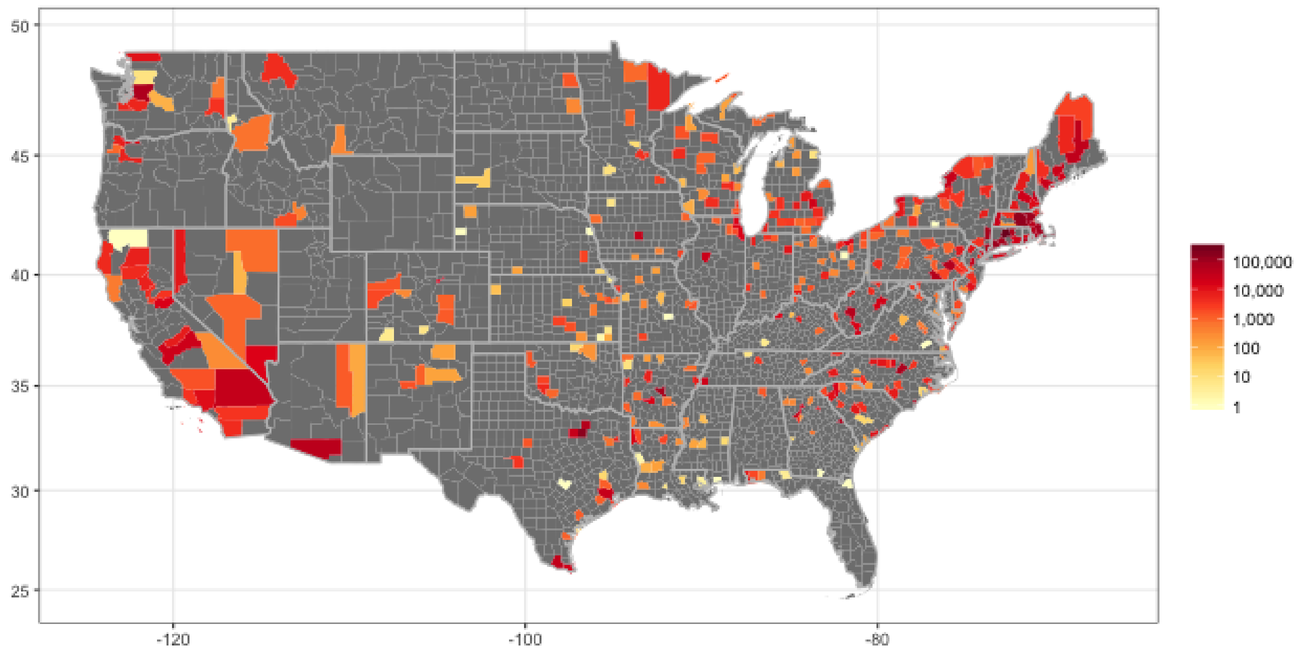

Figure 19. Worst-case scenario (with regressors): the number of outages during the most significant winter storm event in each county from January 2011 to February 2017 (restricted to events where all regressors are available)

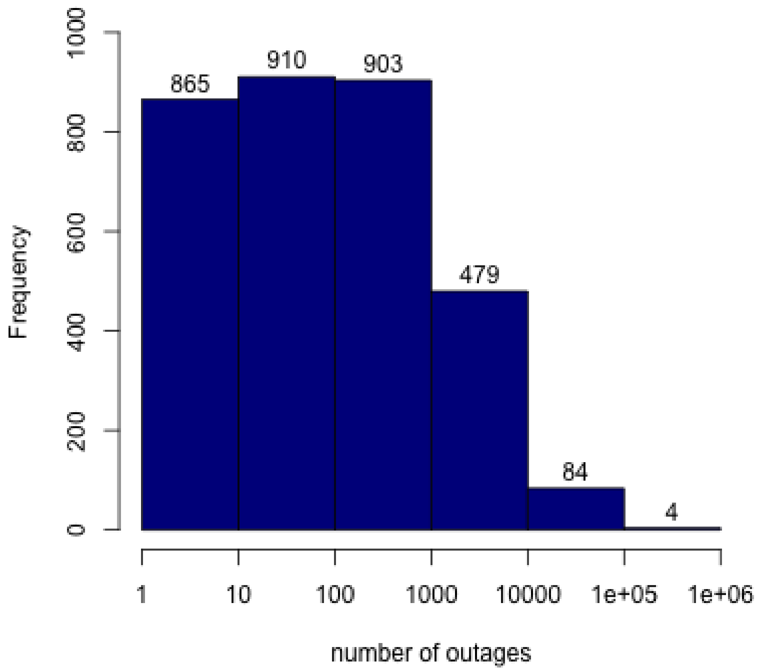

Figure 20. Impact of winter storms: the frequency of reported outages of different sizes

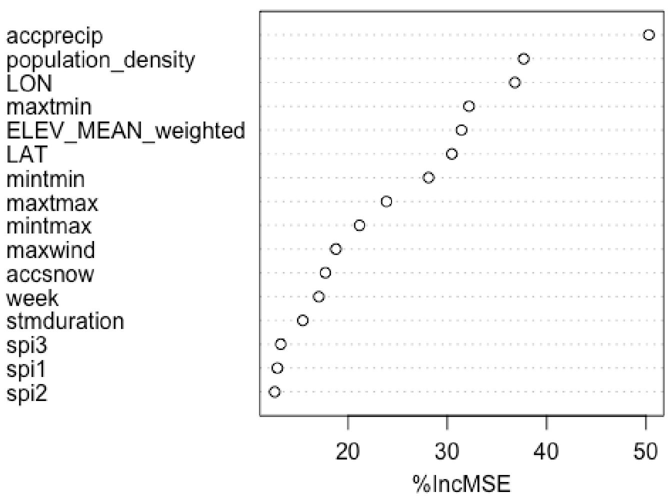

Figure 21. Variable importance plot for the reduced model predicting log outages

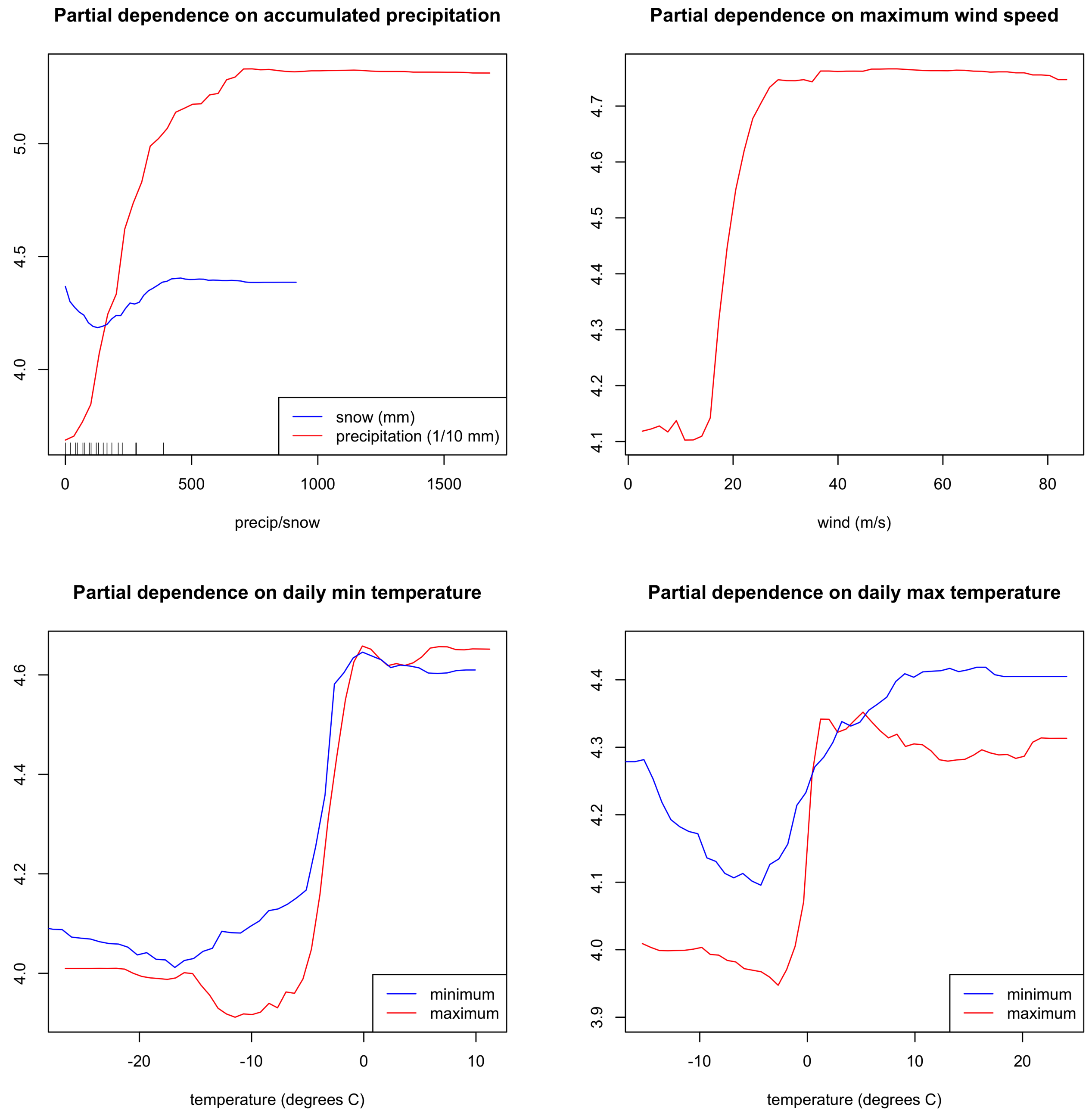

Figure 22. Partial dependence plots for the reduced model predicting log outages

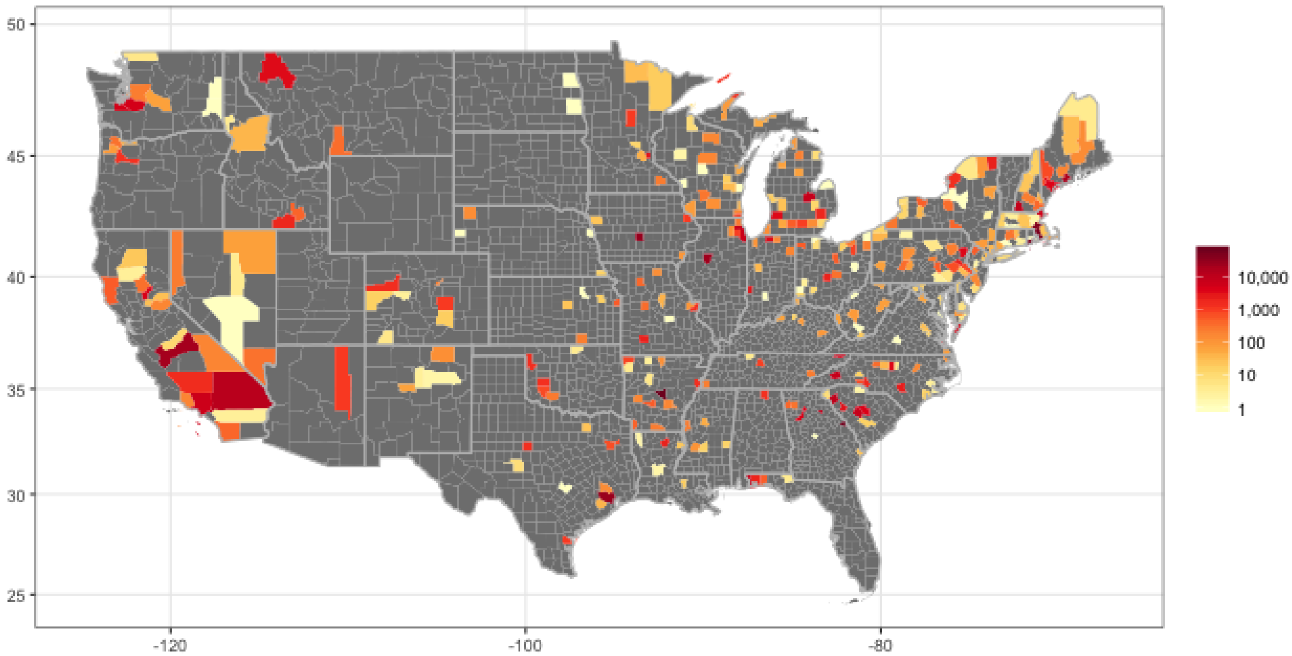

Figure 23. Predicted outages for 10% hold-out sample (at most one from each county)

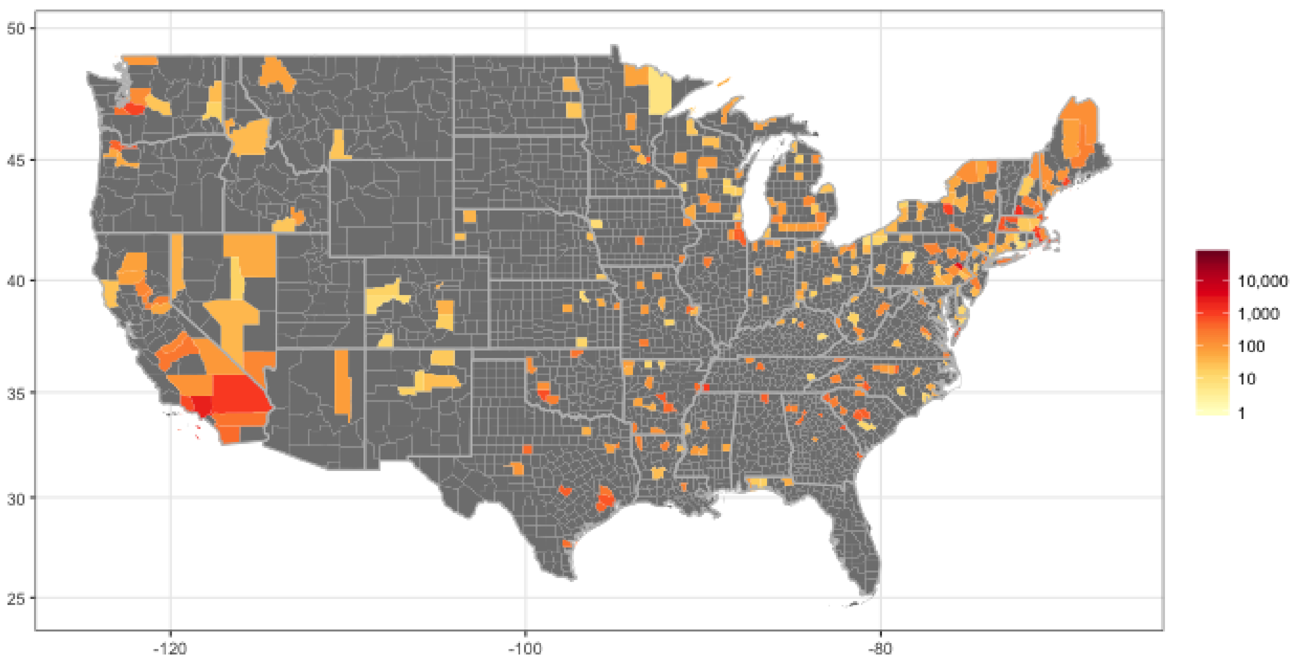

Figure 24. True outages for 10% hold-out sample (at most one from each county)

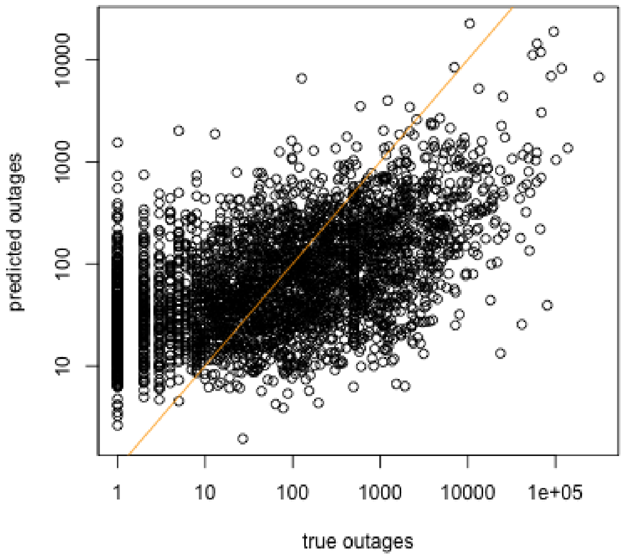

Figure 25. Predicted vs. true outages for all counties in the training data

2 TABLES

2.1 Tropical Storms and Hurricanes

	Model	Raw counts MSE	Raw counts % Var	Log counts MSE	Log counts % Var	Log(100+counts) MSE	Log(100+counts) % Var
All storms	Full	264036997	41.62	3.3395	62.92	0.8989	69.79
	3PC	240426562	46.84	3.2335	64.1	0.8651	70.92
	Reduced	224447152	52.95	3.2057	64.41	0.8481	71.67
Hurricanes only	Full	614461338	42.2	2.6211	69.55	1.1446	74.01
	3PC	546744328	48.57	2.5571	70.3	1.1522	73.84
	Reduced	528233936	53.14	2.6906	69.25	1.2136	73.01

Table 2. Comparison of random forest models to predict outages, log outages, and log(100+counts) outages for all storm data and hurricane-only data

	Model	MAE (sd)	RMSE (sd)
All storms	Full	1.49 (0.05)	1.84 (0.05)
	3PC	1.47 (0.04)	1.81 (0.05)
	Reduced	1.45 (0.07)	1.79 (0.07)
Hurricanes only	Full	1.28 (0.07)	1.62 (0.09)
	3PC	1.45 (0.06)	1.79 (0.07)
	Reduced	1.49 (0.05)	1.84 (0.05)

Table 3. Cross-validation results of random forest models to predict log outages for all storm data and hurricane-only data

	Model	MAE (sd)	RMSE (sd)
All storms	Full	0.725 (0.02)	0.947 (0.03)
	3PC	0.714 (0.03)	0.935 (0.04)
	Reduced	0.702 (0.02)	0.923 (0.04)
Hurricanes only	Full	0.879 (0.03)	1.089 (0.04)
	3PC	0.869 (0.06)	1.081 (0.07)
	Reduced	0.885 (0.08)	1.112 (0.08)

Table 4. Cross-validation results of random forest models to predict log(100+counts) outages for all storm data and hurricane-only data

RF Model	Predicted count interval	Bias	St.Dev	RMSE	n
	0-100	-0.0964	1.0185	1.0228	1861
Log10(counts)	>100-1000	0.7379	1.5377	1.7050	1000
	>1000-10000	0.6762	0.9303	1.1493	495
	>10000-100000	0.7917	0.9597	1.2420	180
	>100000	0.7487	0.5466	0.9122	11
	0-100	-0.6279	0.4336	0.7625	1555
Log10(100+counts)	>100-1000	0.1086	0.7005	0.7086	1241
	>1000-10000	0.4457	0.7270	0.8522	560
	>10000-100000	0.6379	0.7214	0.9614	177
	>100000	1.0050	0.7738	1.2512	14

Table 5. Bias, standard error, and root mean square error comparison for predicting all the storm data

RF Model	Predicted count interval	Bias	St.Dev	RMSE	n
	0-1000	-0.0448	0.9016	1.004	154
Hurricane Matthew	>1000-10000	0.4110	1.255	1.313	81
	>10000-100000	1.641	1.839	2.448	40
	0-1000	-0.4817	0.5481	0.7259	54
Hurricane Sandy	>1000-10000	0.8232	2.321	2.457	207
	>10000-100000	2.555	3.1618	4.0433	57
	0-1000	-0.0943	1.2801	1.2799	174
Hurricane Irene	>1000-10000	1.3130	4.3249	4.503	124
	>10000-100000	2.0168	2.2807	3.0296	57

Table 6. Bias, standard error, and root mean square error comparison for predicting different hurricanes

2.2 Winter Storms

	Raw counts		Log counts	
Model	MSE	% Var	MSE	% Var
Full	65880031	13.67	5.1875	26.73
5PC	63645150	16.60	5.1638	27.06
Reduced	62687074	17.85	5.1503	27.25

Table 7. Comparison of random forest models to predict outages and log outages during winter storms

REFERENCES

[1]Anderson, B.and Schumacher, A., Crosson, W., Al-Hamdan, M., Y., M., Ferreri, J., Chen, Z., Quiring, S., and Guikema, S. (2017). Hurricaneexposuredata: Data characterizing exposure to hurricanes in united

Model	MAE (sd)	RMSE (sd)
Full	1.88 (0.09)	2.30 (0.10)
5PC	1.88 (0.10)	2.29 (0.11)
Reduced	1.88 (0.09)	2.29 (0.10)

Table 8. Cross validation results of random forest models to predict outages and log outages during winter storms

states counties. *R package version 0.0.1*.

[2]Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[EAGLE-I]EAGLE-I. Energy infrastructure mapping system and the real time energy monitoring dashboard. https://eagle-i.doe.gov. 2017-04-18.

[4]Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC press.

[5]Liu, H., Davidson, R. A., and Apanasovich, T. V. (2008). Spatial generalized linear mixed models of electric power outages due to hurricanes and ice storms. *Reliability Engineering & System Safety*, 93(6):897–912.

[6]McKee, T., Doesken, N., and Kleist, J. (1993). The relationship of drought frequency and duration to time scales. *Eighth Conference on Applied Climatology*, (January):179–184.

[7]Nateghi, R., Guikema, S., and Quiring, S. M. (2014a). Power outage estimation for tropical cyclones: improved accuracy with simpler models. *Risk analysis*, 34(6):1069–1078.

[8]Nateghi, R., Guikema, S. D., and Quiring, S. M. (2014b). Forecasting hurricane-induced power outage durations. *Natural hazards*, 74(3):1795–1811.

[NOAA National Severe Storms Laboratory]NOAA National Severe Storms Laboratory. NWS Storm Events Database. `https://www.ncdc.noaa.gov/stormevents`. 2017-06-15.

[10]Quiring, S. M., Schumacher, A. B., and Guikema, S. D. (2014). Incorporating hurricane forecast uncertainty into a decision-support application for power outage modeling. *Bulletin of the American Meteorological Society*, 95(1):47–58.

[11]Quiring, S. M., Zhu, L., and Guikema, S. D. (2011). Importance of soil and elevation characteristics for modeling hurricane-induced power outages. *Natural hazards*, 58(1):365–390.